Extração de Conhecimento de Bases de Dados Biológicas

TRABALHO

O trabalho consta da análise de um conjunto de dados usando o programa R e os packages do Bioconductor (em grupos de 3 elementos). A informação deverá ser recolhida num relatório a ser enviado até ao dia **19 de junho de 2015**.

Os dados a analisar deverão ser provenientes de um conjunto de dados de expressão genética, que poderá escolher de entre dados publicados na literatura ou procurar em bases de dados apropriadas (e.g. GEO do NCBI ou ArrayExpress do EBI). Serão fornecidas pelos docentes algumas sugestões de possíveis conjuntos de dados a analisar. A escolha do conjunto de dados deverá ser validada pelo docente até ao dia 5 de junho.

Deverá desenvolver scripts em R/ Bioconductor para carregar os dados, pré-processá-los/ fazer a sua filtragem (se necessário) e fazer o conjunto de análises que lhe pareçam apropriadas. Deverá incluir, como requisito mínimo dois dos três tipos de análise:

- análise de expressão diferencial (e opcionalmente análise de enriquecimento)
- clustering de genes e/ou amostras
- análise preditiva (e.g. classificação de amostras)

A apresentação dos resultados deverá incluir:

- explicação dos dados, sua origem e relevância
- apresentação do pipeline de análise e ferramentas usadas
- apresentação dos principais resultados obtidos
- discussão (podendo relacionar-se com a publicação de onde os dados originaram)

Deverá submeter no sistema de e-learning o relatório em formato HTML ou PDF e ainda um ficheiro com os comandos R (extensão Rmd) que realizou de forma a que a análise se possa reproduzir. Para a geração do relatório deverá usar o *R markdown* que lhe permite incluir os comandos em R utilizados e mostrar os respetivos resultados, introduzindo texto para a descrição e discussão dos resultados, gerando um relatório em HTML ou PDF. Como sugestão, use o programa Rstudio que inclui as funcionalidades necessárias para a utilização do R markdown.